DOCUMENT RESUME

ED 409 373 TM 026 943

AUTHOR Jakwerth, Pamela M.; And Others
TITLE Validity in Cross-National Assessments: Pitfalls and Possibilities.
PUB DATE Mar 97
NOTE 37p.; Papers presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Collected Works - General (020) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Cross Cultural Studies; *Curriculum; Educational Assessment; Elementary Secondary Education; Evaluation Methods; Foreign Countries; *International Education; Meta Analysis; *Test Validity
IDENTIFIERS Domain Knowledge; Opportunity to Learn; *Third International Mathematics and Science Study

ABSTRACT

Five interrelated papers that explore the theme of validity within the context of cross-national assessments are grouped in this collection. These papers were presented at a symposium at the annual meeting of the American Educational Research Association. These papers draw on data from the Third International Mathematics and Science Study (TIMSS). The TIMSS involved more than 12,650 schools, 25,300 teachers, and 655,000 students in about 50 countries. A central problem in the evaluation of results of cross-national assessments is that of content validity as evaluated in relation to the specific domain to which test scores are intended to relate. In international assessments, as in national studies, more than one type of domain is of potential interest. Some domains relate to the explicit and implicit intended goals of the curriculum, the "intended" curriculum, and others related to what is actually implemented in the classroom, the implemented curriculum. The degree to which a cross-national assessment reflects a country's curriculum and instruction has great impact on the interpretation of results. The following papers are included: (1) "Opportunity To Learn and the Pitfalls of International Rankings: A Validity Issue?"; (2) "Domain Definitions for Curriculum-Sensitive Tests: Improving the Content Validity of Cross-National Assessments"; (3) "Evaluating Test-to-Curriculum Match: Indices of Content Validity for Curriculum-Sensitive Assessment"; (4) "Item-Topic Clusters, Disaggregation, and Variety of Statistics: Some Approaches to Solving the Validity Dilemma in Cross-National Assessments"; and (5) "Validity Issues in Cross-national Relational Analyses: A Meta-Analytic Approach to Perceived Gender." References, when included, follow the individual papers. (SLD)

Validity in Cross-National Assessments: Pitfalls and Possibilities

Pamela M. Jakwerth
Leonard J. Bianchi
Richard T. Houang
William H. Schmidt
Gilbert A. Valverde
Richard G. Wolfe
Wen-Ling Yang

This symposium will explore important issues concerning validity in cross-national educational assessment. Issues, dilemmas and possible solutions will be illustrated drawing upon data from the Third International Mathematics and Science Study (TIMSS)[1]. Five interrelated papers will be presented, each of these designed to complement each other and explore the theme of validity within the context of cross-national assessments.

The problem

For more than three decades, there have been several cross-national assessments of educational achievement. All have claimed, to a greater and lesser degree, to provide important data useful for ascertaining the effectiveness of national educational systems[2]. However, almost since their inception, such studies have been challenged on a variety of grounds. One particular challenge has been the implicit assumption that the student achievement scores reported in such studies can in fact be attributed to their educational experiences, and thus represent a valid assessment of the comparative effectiveness of national educational systems. At the heart of this challenge lies the question of whether or not these assessments do in fact measure the domains that they claim to, and whether or not the processes that explain the performance of students in the test are actually related to the educational experiences of the students (Airasian and Madaus, 1983).

The problem confronted is one of content validity, which is evaluated in relation to the specific domain to which test scores are intended to relate (Crocker, Miller, & Franks, 1989; Messick, 1989). During the construction of an assessment, items should be written so that they adequately sample the specified domain (Sireci, 1990). The theory being that the more representative the items are to the domain of interest, the greater the chances that the examinees' performance on the sample of items will mirror their performance within the entire domain (Messick, 1989).

[1]A recently completed study of mathematics and science education conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) in approximately fifty nations, involving more than 12,650 schools, 25,300 teachers and 655,000 students. It involved the comprehensive study of curricula (course-taking, textbooks, curriculum guides), teaching (opportunity to learn, instructional practices, teacher background, and teachers' goals for content coverage), and student achievement.

[2]See, for example: Foshay, A.W. (ed) (1962) Educational Achievement of 13 year olds in Twelve Countries, Hamburg: UNESCO Institute of Education; NCES. (1985) International Education Statistics, Summary of Discussions, Education Indicators Conference, April 11-12, 1985, Washington D.C.: U.S. Department of Education, National Center for Education Statistics; and Husen, Torsten (1992) "Policy Impact of IEA Research". in R.F. Arnove, P.G. Altbach and G.P. Kelly (eds) Emergent Issues in Education: Comparative Perspectives, Albany: State University of New York Press.

In international assessments (and national assessments as well), more than one primary type of domain is of potential interest (Schmidt & McKnight, 1995). Some of these domains relate to the explicit and implicit *intentional* goals of an educational system (Burstein and others, 1990), often called the intended curriculum. Other domains relate to what is actually implemented in the classroom. Domains of the first type are specified in formal statements of curricular goals and objectives, or in textbooks and other instructional materials. The second type of domain might be termed (after Schmidt, 1983), the *instructional domain*, also known as the implemented curriculum.

Thus it is crucial to determine the domain of interest, that is, whether measures of achievement should reflect what students are intended to learn, the content of their textbooks, what they are taught, what most nations achieve, or something else. The degree to which cross-national assessments reflect a country's curriculum and/or instruction impacts the interpretation of results (Guiton & Oakes, 1995; Linn & Baker, 1995).

These questions of content validity still beset international assessment and have, at best, been unsatisfactorily addressed in most recent studies. Additionally, by-state comparisons in the reporting of national assessment results confront similar dilemmas.

This symposium

Presenters in this symposium have all been involved in data analysis and writing of recently published books reporting international and national results of TIMSS, undoubtedly the most ambitious cross-national assessment attempted this decade. These books have been published and presented to the public at national and international press conferences. Each paper presents a series of problems associated with validity in the context of international assessments, and illustrates these with data from the TIMSS, they also use these data to illustrate methodologies with the potential to address and resolve some of these dilemmas. The symposium thus intends to share with the AERA community a series of methodologies that address validity concerns in their attempt to advance the state of the art in cross-national assessment.

Opportunity to Learn and the Pitfalls of International Rankings: A Validity Issue?

Introduction

Studies comparing the structure of educational systems and the performance of students in nations across the world have been a reality for over 30 years. Educators, policy-makers, and researchers maintain that comparative cross-national studies provide nations with a broad perspective for ascertaining the effectiveness of their educational systems (Linn & Baker, 1995; Mislevy, 1995; Porter, 1990; Robitaille, McKnight, Schmidt, Britton, Raizen & Nicol, 1993; Schmidt & Valverde, 1995). Information from these studies can be used as input for policy decisions aimed at educational improvement. Comparative studies also are conducted *within* nations to monitor educational effectiveness. Within the United States, for example, such studies may use results of student achievement testing to compare states (e.g., National Assessment of Educational Progress - NAEP), districts (e.g., Michigan Educational Assessment Program, MEAP; California Learning Assessment System, CLAS; Kentucky Instructional Results Information System, KIRIS), or programs within districts (LaPointe, 1991).

Researchers conducting comparative-education studies typically collect a wide array of information from participating educational systems. In addition to collecting student performance data, comparative researchers may collect descriptive information related to the structure and processes of each educational system or attitudinal information from stakeholders such as students, teachers, or administrators. Despite the availability of descriptive information, however, the public, educators, and policy-makers focus much of their attention on student performance results, and, often, these results receive the primary emphasis in reporting and analysis (Husen, 1987; Linn, 1988).

One popular approach for reporting student performance results in cross-national studies is to rank countries using total scores, or selected sub-scores, on tests presumed to measure student achievement in various subject areas. The process is simple. Batteries of test items are administered

to carefully selected samples of students in selected grade or age levels and in selected countries. Then, subject-matter proficiency is estimated by producing a score (e.g, percent of items passed, sum of items, scale scores based on a Rasch single-parameter IRT model - such as was used in TIMSS). Countries are then ranked according to these "scores," and although actual performance results are reported along with the rankings, it is the rankings that seem to have the most inherent meaning. The common interpretation of these rankings is that students in nations ranking at or near the "top" are achieving, or have learned, more than students in nations ranking lower. The implication is that the nations at the top have more effective educational systems, at least in particular subject areas, than do the nations at the bottom.

An example of this comes from the Third International Mathematics and Science Study. A portion of the data from this study was recently released. Despite many efforts to de-emphasize a focus on rankings in these results, country rankings were reported and have been widely discussed in the press, among policy makers, and among researchers. The important messages in TIMSS, like those messages from many previous cross-national studies, are in danger of being lost in what has become an "international horse race."

Is this "horse race" really all that bad? What are the problems with using a single score for cross-national comparisons? The problem is that when the results show students in country A did better than students in country B, does it mean students in country A are smarter, are better "equipped" to take the test, or have more exposure to the content areas covered by the test? In the narrowest sense, test scores provide an indication of how well students perform with respect to the collection of items they were administered. The uncertainty is in what extent we can generalize a student's (or country's) performance beyond the items on the test battery. Answers to questions about the meaning behind interpretations of subsequent country ranks on these scores depend on the ability of the test that was used to obtain the rankings to measure what it was intended to measure

(i.e., the validity of the test) and, sometimes, determining exactly what a particular study was intending to measure in the first place.

The Focus of Cross-National Achievement Studies

Often there are conflicting purposes for conducting cross-national comparative-achievement studies. Policy makers may be interested only in student achievement comparisons in and of themselves. However, most cross-national studies typically have a purpose beyond merely ranking countries on student-test performance (Burstein, 1993; Husen, 1983; Postlethwaite, 1987). Some researchers (Bracey, 1991; Burstein, 1993; Linn & Baker, 1995) find it valuable to know how students within a nation perform on test content that is unique to their particular educational system and to compare this performance to the performance of students in other nations on content that is unique to their system. These researchers are less interested in performance differences due to "student attributes" (Burstein, 1991, p. 50) or ability than they are interested in detecting differences due to schooling and determining how and why these differences arise (Burstein, 1991; Husen, 1983). Simply finding out that the students of one nation perform better on a set of items than do students of another nation is not meaningful to educational improvement if student performance cannot be not linked to some characteristic of a particular educational system.

Burstein (1993), in the prologue to his edited volume on SIMS results, recounts the historical purpose behind IEA testing. In it, he quotes from Husen's preface to the 1967 volume on the First International Mathematics Study:

> ...the overall aim is, with the aid of psychometric techniques, to compare outcomes in different educational systems. The fact that these comparisons are cross-national should not be taken as an indication that the primary interest was, for instance, national means and dispersions in school achievements at certain age or school levels. ...the main objective of the study is to investigate the "outcomes" of various school systems by relating as many as possible of the relevant input variables (to the extent that they could be assessed) to the output assessed by international test instruments...In discussions at an early stage in the project, education was considered as a part of a larger social-political-philosophical system. In most countries, rapid changes are occurring...Any fruitful comparison must take account of how education

> responded to changes in the society. One aim of this project is to study how mathematics teaching and learning have been influenced by such development.(p. 30) ...The IEA study was not designed to compare countries; needless to say, it is not to be conceived of as an "international contest" ...its main objective is to test hypotheses which have been advanced within a framework of comparative thinking in education. Many of the hypotheses cannot be tested unless one takes into consideration cross-national differences related to the various school systems operating within the countries participating in this investigation. (in Burstein, 1993, p. xxxii)

Therefore, the value of many comparative achievement studies depends upon the extent to which student test performance reflects achievement that can be attributed to the student's educational experiences (Airasian & Madaus, 1983; Linn, 1987; Mislevy, 1995; Nitko, 1989; Schmidt & McKnight, 1995). Often researchers look to the curriculum of a nation as one indication of these educational experiences, and many comparative studies focus on the success with which educational systems impart to their students a certain defined curriculum. The tests developed for these studies are designed to measure student attainment of this curriculum. A key component to evaluating the validity of these tests is determining how representative the test content is of the corresponding curriculum. Often, measurement specialists refer to this particular component of validity as *content validity.*

Critics of cross-national achievement studies often argue that the tests used in these studies provide, at best, an abstract definition of achievement in a particular subject area and may not adequately represent the curriculum of any participating nation (Linn & Baker, 1995; Mislevy, 1995; Porter, 1990; Westbury, 1992, 1993). The accuracy and meaningfulness of interpretations of cross-national achievement results are impacted by the degree to which the test used in a particular cross-national study, reflects the curriculum of each country in the study (Guiton & Oakes, 1995; Linn & Baker, 1995; McDonnell, 1995; Romberg & Wilson, 1992). Performance results on a test that is not based on a clearly defined domain provides little more than the knowledge of who outperforms who on a specific set of items (Airasian & Madaus, 1983; Robitaille et al., 1993). Interpretations of educational effectiveness or explanations of cross-national differences that are based on such results

are questionable, if not invalid (Airasian & Madaus, 1983; Berliner, 1993; Guiton & Oakes, 1995; Guskey & Kifer, 1990; McDonnell, 1995; Stedman, 1994; Westbury, 1992, 1993).

Evidence of the Need for Caution in the Interpretation of Rankings

If results on the TIMSS achievement test is taken as a measure of student achievement of a particular curriculum, data presented here and in other studies point out the need for more information before meaningful conclusions can be reached from these results. Particularly, these data will highlight the variability in curricula across the world and the subsequent difficulty in a) developing tests that adequately reflect these curricular variations and b) interpreting results from tests that do not. The overarching issue is one of content validity.

The first point will be illustrated here:

- A major component of TIMSS was an extensive, multi-national curriculum analysis (e.g., see Schmidt, McKnight, & Raizen, 1997; Schmidt, McKnight, Valverde, Houang, & Wiley, 1997; Schmidt, Raizen, Britton, Bianchi, & Wolfe, 1997). The analysis yielded data on the mathematics and science curricula of approximately 50 nations. Textbooks were analyzed, curriculum guides were reviewed, and curriculum experts were consulted. The data show extensive variation in the mathematics and science curricula across the world at any given grade level. Variations exist in the number of topics countries intend for inclusion in instruction, the relative emphasis given topics, and the amount of time topics remain in the curriculum. Countries do not expect mastery of the same topics, at the same times, and in the same ways. This variation illustrates the need for caution when interpreting international rankings. It is difficult to identify "an eighth grade curriculum" and then develop a test to measure that curriculum.
- Preliminary data from the curriculum analysis were used when developing the TIMSS achievement test. However, due to the tremendous curricular variability across nations and the desire to over-sample some topic areas, the TIMSS test varied in its match to any particular

curriculum. National Research Coordinators within each country were asked to select the items they felt were appropriate for inclusion on a test of the students in their country (Beaton, Martin, Mullis, Gonzales, Kelly, Smith, 1996; Beaton, Martin, Mullis, Gonzales, Smith, Kelly, 1996). The mathematics achievement test for 13 year old students had 162 items and the science test had 146 items. The number of items chosen for inclusion on country-specific tests ranged from 76 to 162 for tests of mathematics for students at the upper grade and from 59 to 162 for the lower grade. In science the number of items chosen ranged from 58 to 146 for the upper grade and from 20 to 146 for the lower grade. Countries did not choose the same number of items for each grade. Clearly variability exists in what countries deem as appropriate for testing. One must wonder how fair it is to compare countries on tests composed of items that they themselves would not have included without some additional explanation.

- When comparing the curriculum evidence in the aggregate we have found achievement patterns that match patterns of curriculum coverage. Average difference between the lower and upper grade performance was highest on those topics most emphasized across countries in the upper grade for 13-year old students. In mathematics, the top five topics most emphasized in textbooks were equations and formulas; polygons and circles; 3D geometry; 2D geometry; and perimeter, area, and volume. Four of the top five topics in achievement difference between the lower and upper grade students were equations and formulas; polygons and circles; perimeter, area, and volume; and 3D geometry. Less difference is seen in 2D geometry, perhaps because this is not a new topic for students of this age. Additionally, a large difference is seen in congruence and similarity which is not highly emphasized in textbooks. However, many items measuring congruence and similarity also measure polygons and circles, and the amount of textbook space needed to convey the concepts of congruence and similarity are probably minimal. Combining the topic of congruence and similarity with polygons and circles changes the combined rank in

textbooks to second. Additionally in science, which is more difficult to evaluate because of the number of topics and lack of dependence among topics as compared to math, similar patterns are seen. Performance difference is highest in the physical sciences followed by life and earth sciences. This concurs with the ordering in textbooks. Similar conclusions can be reached when comparing teacher coverage with differences in student performance.

These points illustrate the need to understand the curricular variability across nations and the sensitivity of the test to that variability before attempting to reach conclusions based on country ranks.

The Impact of Low Content Validity

Considerable disagreement exists as to the impact of the lack of fit between a test and a domain. One impact of the lack of fit is the perceived importance of the test to stakeholders. Linn (1987) stated, "If a test does not measure the outcomes that correspond to important program goals, the evaluation will surely be considered unfair" (p. 6), especially if it better measures the goals of another program in the study.

Studies have shown that results on tests not well-matched to a domain can be misleading (Berliner, 1993; Linn, 1988; Stedman, 1994; Westbury, 1992, 1993). Others have found that ranks on total scores are unstable, may result in unfair comparisons (Guskey & Kifer, 1991; Linn, 1987; Mislevy, 1995), and are dependent on the relative weighting of sub-topic areas (Cronbach, 1971). IEA studies introduced the notion of opportunity to learn (OTL) as a means of ensuring the technical validity of their findings (McDonnel, 1995). Researchers have shown that opportunity-to-learn the skills being tested is a significant explanatory variable of student performance (Berliner, 1993; Burstein, 1993; Burstein et al., 1990; Husen, 1983; Kupermintz et al., 1995; McDonnell, 1995; Muthen, Huang, Jo, Khoo, Goff, Novak, & Shi, 1995; Purves, 1987; Walker & Schaffarzick, 1974).

Additionally, Westbury (1993) found that differences between the scores of American and Japanese students on SIMS decreased when controlling for curriculum. Studies by Raizen and Jones (1985) found a correlation between mathematics achievement and the number of math courses students take. One particular critic of cross-national studies has stated

> We make curricular decisions different from those that other countries make. Thus differences in achievement are most parsimoniously explained as differences in national curricula, rather than differences in the efficiency or effectiveness of a particular national system of education. (Berliner , 1993, p.),

Differing opinions about the impact of curriculum on student achievement also exist. In a reanalysis of the Westbury data, Baker (1993) still found large differences between American and Japanese scores even when accounting for opportunity to learn. Furthermore, although he did find some curricular impact on test results, Stedman (1994) found that curriculum was just one of many variables having an impact. Phillips and Mehrens (1988) maintained that studies comparing test-to-curriculum match "have not provided any evidence regarding the impact of the mismatch" (p. 34). Mehrens (1984), Mehrens and Phillips (1987), and Phillips and Mehrens (1988) felt that impact of mismatch on achievement would be minimal in norm-referenced testing situations where the curriculum is basically homogenous. However, they surmised that the results could be quite different if comparing "two totally different curricula" (Mehrens & Phillips, 1987, p. 368) or when comparing "countries in which textbooks are not as homogeneous as those in the United States" (Phillips & Mehrens, 1988, p.50).

> It is reasonable to assume that the more different the curricula, the more likely those differences will have an impact on the test scores. Thus if differences in curricula between, for example, the United States and Japan are great, those differences may indeed impact scores on a common test. Examining score differences across countries, we could make incorrect inferences about the quality of the instruction or the quality of the students rather than making correct inferences about the impact of curricular differences on test scores. (Mehrens & Phillips, 1987, p. 358)

Some evidence of this is seen when evaluating country performance on the sub-scales within mathematics and science. The number of countries with performance significantly higher and lower than US students changes across sub-scales. In math, the number of countries outperforming the US on the total score is20. In the sub-scales, this ranges from 9 in data representation and analysis to 30 in measurement. In science, the number of countries outperforming the US on the total score is nine. In the sub-sales this ranges from 1 in environment to 13 in physics. More evidence of this variation will be presented later.

Methodological Issues

Evaluating the content validity of large-scale, comprehensive measures is not easy. Several issues need to be addressed. The first issue relates to the complexity of the curricula and the difficulty in describing it well enough to develop test specifications. Some method for reliably and validly "measuring" the curricula must be used and then rules are needed for turning these measures into test specifications. The first paper will discuss this in more detail. Additionally, curriculum is more than topics; it also includes what students are expected to do with the topics.

A second issue relates to the evidence used to indicate validity or invalidity. Two primary methods exist for evaluating content validity. One method evaluates the "match" between test specifications and a domain; the other compares the performance results of groups of students. These will also be discussed in more detail.

Finally, decisions are made throughout the entire reporting and analysis stage as to the levels to which data will be aggregated (e.g., item vs. total score, individual vs. group) and the test statistics that will be used. These decisions have an impact on the validity of the findings. Various approaches for reporting and analyzing data from cross-national studies will be discussed.

Domain Definitions for Curriculum-Sensitive Tests: Improving the Content Validity of Cross-National Assessments

Introduction and Focus Paper

Results from cross-national studies of student achievement, such as the recent Third International Mathematics and Science Study, are often tainted with controversy. These studies are hailed by many as among the most important in education, and they often have a major impact on policy decisions for years after they are conducted. At the same time, critics of these studies warn the public that their results are not to be taken seriously as they are invalid, and, therefore, practically meaningless. Among the most serious criticisms of the validity of cross-national achievement studies are those related to the differing curricula of the educational systems involved in the studies and the problems that arise in test development and reporting as a result of these curricular differences (Berliner, 1993; Linn & Baker, 1995; Stedman, 1994; Westbury, 1992, 1993).

IEA forefather Torsten Husen once stated that, "comparing the outcomes of learning in different countries is in several respects an exercise in comparing the incomparable" (1983, p. 455). The difficulty stems from the fact that educational systems are unique to the culture of each country (Passow, 1984; Purves, 1987). They are based upon differing views of development and childhood (Berliner, 1993). They have differing goals which reflect differing social, political, economic, and resource needs and priorities (Schmidt & McKnight, 1995; Schmidt & Valverde, 1995). The time available for formal education in each country is limited, making it impossible to teach everything, and it is highly unlikely that different nations will choose to fill this limited time in exactly the same way (Schmidt & McKnight, 1995).

The variability in curricular goals and offerings across differing educational systems has an impact on the interpretation of results from comparative studies of these systems (Berliner, 1993; Linn & Baker, 1995; Mislevy 1995; Stedman, 1994; Westbury, 1992, 1993). More specifically, the accuracy and meaningfulness (i.e., validity) of the interpretations relate to the degree to which the

test used in a particular cross-national study, reflects the curriculum of each country in the study (Guiton & Oakes, 1995; Linn & Baker, 1995; McDonnell, 1995; Romberg & Wilson, 1992).

One difficulty in developing cross-national assessments with adequate content validity is that different curricula (i.e., domains), or components of a curriculum, may be of interest to educators and researchers who conduct cross-national studies (Schmidt & McKnight, 1995). For example, aside from the particular subject matter of interest, researchers may be interested in the curriculum as laid out in official documents (e.g., curriculum guides, national goals statements) or as laid out in textbooks and other instructional materials. Additionally, some researchers may be interested in the curriculum that is actually delivered by teachers. A crucial, and often ignored, issue in the development of cross-national achievement tests is determining what specific component of a curriculum (i.e., domain) is of particular interest (Airasian & Madaus, 1983; Mislevy, 1995) and, therefore, whether achievement results should reflect what students are intended to learn, what is in text books, what is delivered in the classroom, what the students of most nations achieve, or something else (Airasian & Madaus, 1983). Even when a specific domain is identified, cross-national researchers still face challenges in writing test specifications for that domain. For example, a test could consist of only those topics that all countries include in their curriculum, topics that most countries include in their curriculum, or all topics included in the curriculum of any country (Linn, 1988; Linn & Baker, 1995; Porter, 1990).

Generally, however, cross-national achievement tests are comprised of items that represent an internationally negotiated set of content (Linn & Baker, 1995). Critics of cross-national achievement studies often argue that the tests used in these studies provide, at best, an abstract definition of achievement in a particular subject area and may not adequately represent the curriculum of any participating nation (Linn & Baker, 1995; Mislevy, 1995; Porter, 1990; Westbury, 1992, 1993). Performance results on a test that is not based on a clearly defined domain provide little more than

the knowledge of who outperforms who on a specific set of items (Airasian & Madaus, 1983; Robitaille et al., 1993). Interpretations of educational effectiveness or explanations of cross-national differences that are based on such results, then, are questionable, if not invalid (Airasian & Madaus, 1983; Berliner, 1993; Guiton & Oakes, 1995; Guskey & Kifer, 1990; McDonnell, 1995; Stedman, 1994; Westbury, 1992, 1993).

This paper explores the concept of content validity as applied to tests used to compare student achievement across nations. It begins with a conceptual discussion of validity and moves on to an empirical example of how curriculum data can be used to specify potential domains for cross-national achievement testing. Data from the TIMSS analyses of textbooks, curricula, and teachers' instructional practices are used to explore methods for domain specification based on varying criteria of content validity. The primary purpose was to use the results of an extensive multi-national curriculum analysis to develop several sets of "test-blueprints" based on different methods of summarizing the curriculum data.

Study Design

The Third International Mathematics and Science Study (TIMSS) was a multi-component study of curriculum, instructional practices, and student achievement in mathematics and science in nearly 50 countries. One goal of the study was to explore the effects of content coverage on student achievement. A detailed curriculum framework for mathematics and science was developed in order to facilitate accomplishment of this goal (Robitaille, et al., 1993). The framework provided a hierarchical list of topics (e.g., algebra, earth features) and performance expectations (e.g., knowing, communicating). Which were used to code the content of all curricular materials and instruments used in the study.

I used content data from the TIMSS curriculum analysis and teacher questionnaires to write a series of "test blueprints" based on different methods of content selection. From the curriculum

analysis I used data from three sources in each country: a) the mathematics curriculum guides for 13-year old students, b) the mathematics textbooks for these students, and c) curriculum experts. Each source provided a slightly different type of information. Teachers provided information on the amount of instructional time they devoted to topics. I was interested in the variability of test content across the sources and methods of selecting test content. The questions of interest were

- How much variation in the content of mathematics curricula for 13-year-old students exists across the nations involved in the study? What test specifications provide a good curricular match across the countries?

I reviewed the content of each curriculum source and summarized it across countries and across topics. I compared topic inclusion and coverage both across and within countries.

I wrote test blueprints for three "inclusive" tests (i.e., the same test for each country, combining curriculum information across countries) based on each curriculum-data source using the following methods

1. a *strict intersection* (SI) method that included only the topics in all countries' curriculum within each of the data sources,
2. a *70% intersection* (7I) method that included only the topics common to at least 70% of the countries' curriculum within each of the data sources, and
3. a *union* (UN) method that included all topics in any of the countries' curriculum within each of the data sources.

Conclusions

I set out in this study to develop test blueprints for cross-national assessments that validly measure student achievement of topics in the mathematics curriculum for 13-year-old students. However, the variation within and across nations in curriculum and lack of an adequate item pool complicated this goal. Through my analyses, I found that

1. The mathematics curriculum for 13 year old students (as defined in curriculum materials, by experts, and teachers) varies across nations, and variation also exists across curricular sources within nations. Some countries include few topics in their curricula (as indicated by the data sources), and others include many. Some countries focus on particular topics; others spread their focus across many topics. However, some commonalties do exist, with a handful of topics either missing from most countries' curriculum sources or being highly emphasized in most countries. Variations within each country's data sources point out the need for multiple representations of math curricula.
2. Test blueprints varied according to test purpose. Topic coverage and emphasis were inconsistent across the blueprints due to the variability in the curriculum sources. Some blueprints, though, were very similar to one another (e.g., all the union blueprints), while others were very different (e.g., the strict intersections). Each blueprint provides a different look at student achievement. For example, because the strict-intersection tests do not represent the entire curriculum of any country, the weighting of the topics on the test relative to other topics in a country's curriculum is lost. However, the strict-intersection tests do provide information on how students perform on topics included in the curriculum of all countries. Furthermore, the unique tests written to match each country's curriculum provide an indication on how students performed on those topics that are important within a particular country. However, comparisons of student performance when all students do not take the same test are complicated. Finally, tests based upon teacher coverage data would indicate how students perform on what teachers say they teach.

The study has also demonstrated the importance of the first rule in test development - identify the purpose of the testing. Simply starting with collections of items and piecing them together to fit a content map is not adequate. Test developers need to clearly articulate what they are attempting to measure and what types of inferences are appropriate and inappropriate. Unfortunately, this is often

neglected, and consumers are left to guess at the domain, or researchers imply that the test represents more than it actually does. Secondary analysts may also be guilty of applying test results to too broad a domain. These situations can be avoided if test developers clearly describe the testing domain.

Evaluating Test-to-Curriculum Match: Indices of Content Validity for Curriculum-Sensitive Assessment

Introduction

Content Validity in Cross-National Assessment

One important goal in cross-national studies of educational systems is determining the effects that different curricula have on student achievement. Often these studies will collect an array of data on curriculum materials, teacher coverage of content, and student performance. Systems are compared on overall performance, and analyses are conducted in an attempt to link this performance back to variations in curricular offerings and coverage. A key component to evaluating the validity of assessments used in cross-national studies is determining how representative the test content is of the curriculum of countries involved in a particular study, that is, evaluating the content validity of the assessment.

The content validity of a test is evaluated in relation to the specific domain (in this case, a specific curriculum) about which test scores are used to make inferences (Crocker, Miller, & Franks, 1989; Fitzpatrick, 1983; Messick, 1989). The more representative the items are of the domain of interest, the greater is the chance that student performance on the sample of items will mirror their performance within the entire domain (Messick, 1989). A test may have high (content) validity in relation to one domain but low (content) validity in relation to another, and all persons who use the results of a particular test may not be interested in the same domain.

Critics of cross-national studies often cite the cross-country variations in test-curriculum match as reasons behind the invalidity of the conclusions from such studies. The tests are said to be biased for those countries with which an adequate match does not exist. However, developing a test with an adequate match to the variations in country curricula is not easy. First, substantial curricular variation exists across countries and different methods of summarizing this curricula on a test will

lead to differences in how well the test matches the curriculum of any particular country (Jakwerth, 1996).

Another difficulty stems from the politics of item negotiation. Decisions about the specific content of cross-national achievement tests evolve through years of negotiation. Reaching even a minimal level of consensus from participating nations demands sensitivity to the unique concerns and political realities of each nation. Often, reaching consensus entails cutting corners in test development and adding or deleting certain items or topics despite specifications to the contrary.

Finally inadequacies of the item pool available to test developers hinder the development of tests with maximum content validity. Item writing is an arduous and costly process. It is even more difficult in the cross-national arena as it involves developing items that transcend cultures and translations. Often, researchers will draw from existing item pools when constructing large-scale achievement tests (Garden & Orpwood, 1996; Husen, 1983). However, the existing item pool may not always adequately represent the range of topics and behaviors included in the curricula of all nations. Items, especially those measuring higher-order thinking or complex reasoning, may be sparse, and resources may prohibit the development of enough items to overcome the deficits.

The reality of these constraints may mean that cross-national tests will never allow for a perfect match to all potential curricula. Therefore, researchers must continue to explore ways to use the information available on cross-national curricular differences to aid in the interpretation of cross-national-achievement results (Linn & Baker, 1995; Porter, 1990).

Methods for Evaluating Content Validity

Two primary approaches exist for evaluating test-content validity (Airasian & Madaus, 1983; Leinhardt & Seewald, 1981). The first approach uses test results to compare the performance of individuals who have been exposed to curricular content (or some other variable of interest) with the performance of those who have not been so exposed. The intent is either to determine if test scores

discriminate between these two groups or to find items that do discriminate between the groups (Airasian & Madaus, 1983; Burstein, 1991; Muthen et al. 995). This approach includes the use of IRT, intra-class correlations, factor analysis, and generalizability theory. The methodology is used *post hoc* and does not directly evaluate the content being measured by test items (Airasian & Madaus, 1983).

The second approach to evaluating content validity relies on a judgment of the overlap between a test and a domain (Airasian & Madaus, 1983; Crocker et al., 1989; Leinhardt, 1983; Leinhardt & Seewald, 1981; Messick, 1989). Generally, a taxonomy to which the domain and test are matched is developed (e.g., Burstein, 1986; Gamoran, Porter, Smithson & White, 1996; Schmidt et al., 1983). This taxonomy may include only topics or a matrix of topics and cognitive processes. In some cases (e.g., Leinhardt, 1983; Leinhardt & Seewald, 1981; Schmidt & McKnight, 1995), actual test items are matched to textbooks or teacher coverage.

Focus of the Paper

The purpose of this study was to use the results of an extensive multi-national curriculum analysis to analyze the content of a cross-national mathematics achievement test in relation to the curriculum of nations administering the test. The ultimate goal was to use this information to enhance the validity of cross-national comparisons of student achievement. My primary focus was on the relationship between test items and curriculum as a key element of test validity.

Study Design and Procedures

I compared the mathematics-curriculum data collected through the TIMSS document analyses to the content of the TIMSS mathematics-achievement test for 13-year-old students. I used data on a) intended mathematics topic coverage for 13-year old students as reported by curriculum experts in each country (expert-topic mapping), b) topic inclusion in mathematics curriculum guides for 13-year old students in each country, and c) the proportion of mathematics textbooks for 13-year

old students devoted to each mathematics topic (Schmidt, McKnight, Valverde, et al., 1997). The mathematics topics are identified in curriculum frameworks specifically designed for the TIMSS study (Robitaille et al., 1993). The curriculum frameworks enable one to articulate content in terms of topics and performance expectations. The content of all curriculum materials and the TIMSS achievement test were coded according to the frameworks. My primary question was

- How well does the content of the TIMSS mathematics achievement test for 13 year old students match the curricula of the study countries individually and as a whole?

I evaluated test-curriculum match using several methods. For most analyses, I treated each set of topic proportions (i.e., the proportions of emphasis computed for the expert-topic-mapping- and curriculum-guide-data sources and the proportion of textbook blocks in the textbook-data source) for each country as a different "profile" of the mathematics curriculum for the country. Likewise, topic weights (i.e., proportions of items allocated to each topic) on the achievement instrument provided a "profile" of test emphasis. Thus, I sought to compare the similarity of the three curriculum profiles for each country to or their dissimilarity from the test profile.

I looked at the match between the curriculum profiles and the test profile separately for each country. I conducted six different analyses to estimate test-curriculum match. First, I calculated the proportion of items on the mathematics-achievement instrument that measured topics appearing in each of the three curriculum profiles. Second, I calculated the proportion of each curriculum profile that was tested on the achievement instrument. Third, I calculated differences between measures of topic inclusion (i.e., presence) on the achievement instrument and topic inclusion in each of the four curriculum profiles. Fourth, I calculated differences between topic weights (i.e., the proportion of items for each topic) on the achievement instrument and topic emphasis proportions in each curriculum profile. Finally, I computed correlations and Euclidean-distance measures,

$\sqrt{\sum_{j=1}^{44}(W_j^T - W_{ji}^C)^2}$ – where W_i^T is the weight of topic j on the achievement instrument and W_{ji}^C is the weight of topic j in the curriculum of country i, between the topic weights on the achievement instrument and topic-emphasis proportions in each of the four curriculum profiles.

I also summarized the three curriculum-data sources in three ways. The first was using average proportions for topics included in any countries' data; the second was using average proportions for topics included in 70% of the countries' data; the third was using average proportions for topics included in all countries' data (Jakwerth, 1996). I then repeated the same analyses described above comparing the "profile" of topic weights for each of the nine sets of test specifications with the assessment instrument's "profile" of topic weights.

Summary

I found that the content of the achievement-test instrument is more similar to the content of the curriculum of some countries than others and is more similar to the content of some of the curriculum-data sources than others. Some of this variation is due to topics that were not tested on the test, but were found in the curricula. This "differential match" has implications for the validity of inferences made from the test, but final conclusions about test validity will depend on the purpose for which the test will be used. The impact of the mis-match needs to be balanced with other information provided by the tests. Additionally, each of the indices of content match yielded differing pieces of information. Several of the statistics should be reported.

One recommendation is that a higher quality item pool be developed for cross-national work. Several topics important to many countries were missing from the TIMSS test, and items measuring complex applications of topic knowledge and understanding were not available for all topics. The items were not a comprehensive representation of the performance expectation aspect of the curriculum framework. It is difficult to determine how country performance might vary if more items

measuring higher order skills were included on the test. Many countries expect their students to demonstrate complex use of subject matter. If researchers want to adequately measure such skills, better items will need to be developed. Fortunately, within the U.S. research is being conducted on content standards as well as performance standards (Linn & Baker, 1995). Cross-national researchers should look to these studies to guide their research. Until better item pools are developed, results of cross-national achievement testing should be interpreted with caution.

Finally, my recommendation is that researchers take into account the complexity of the curriculum and items when evaluating test-curriculum match. A clear match with curriculum is unlikely to emerge by focusing only on topics. Two countries may demonstrate the same level of coverage on a topic, but have different expectations for performance. Likewise, two items may measure the same topic, but be very different in the type of performance or application expected. Replications of the analyses in this study may produce different results if performance expectations were included in the analyses.

Item-Topic Clusters, Disaggregation, and A Variety of Statistics: Some Approaches to Solving the Validity Dilemma in Cross-National Assessments

One of the goals of cross-national studies is to enable countries to understand their educational systems and the relationship of systemic characteristics to student achievement. Once an appropriate analysis is performed within a country, countries can be compared for further understanding of the differences across countries in student achievement, and how unique country variables relate to this achievement. This leads to a quest to determine the most appropriate measures to use for capturing country differences.

The purpose of this paper is to explore how the aggregation of data over items and over students as well as the use of different statistics can impact one's search for curricular effects. Specifically, it illustrates the information that is lost as one moves from describing performance on individual items to describing performance across items measuring a variety of topics or as one moves from describing the performance of individual students to describing the performance of all students within a country. The paper also discusses how different statistics (e.g., mean scores, difference scores, variance) may be more sensitive to or more descriptive of curricular effects. Various ways of reporting and using the data from cross-national achievement studies will be explored and their relationship to validity will be discussed.

Item-Topic Clusters and Performance Variation

The increasing complexity of subject matter calls into question the unidimensionality of test domains. Lack of unidimensionality raises questions about the meaning of total scores used in country ranks and subsequent analyses (Airasian & Madaus, 1983; Maeroff, 1983). Researchers (Burstein, 1991; Kupermintz, Ennis, Hamilton, Talbert, & Snow, 1995; Maeroff, 1983; Muthen et

al., 1995) have suggested that mathematics scores aggregated over different topics represent general-math ability rather than math achievement that can be linked to curriculum or instruction. Student performance varies, sometimes significantly, across sub-topics (Ariasian & Madaus, 1983). This general-math factor may be so strong that it masks any correlation between curriculum and achievement (Burstein, 1991). Better linkage between tests and curriculum is obtained at the sub-topic level (Airasian & Madaus, 1983; Burstein 1991; Mislevy, 1995); although, some researchers suggest that the most useful performance results are at the item level (Guskey & Kifer, 1990; Mislevy, 1995). As Mislevy (1995) has stated, "The outcome for every individual task in an international assessment tells a story in its own right. Assessments with hundreds of tasks, like those of IEA and IAEP, tell hundreds of stories" (p. 426).

Several researchers (Beaton, Martin, Mullis, Gonzales, Kelly, & Smith, 1996; Beaton, Martin, Mullis, Gonzales, Smith, & Kelly, 1996; Burstein, 1993; Jakwerth, 1996) have attempted to study the impact that altering test content has on performance results. When scores and ranks are compared on the total tests, little difference is seen in rank ordering of countries. However, some difference is seen in score distributions.

More variation is evident within countries, however, when looking at individual topic scores. Mathematics items were divided into 20 topics according to framework codes, and science items were divided into 17 topics. The number of items within topics ranged from between 5 and 28. In mathematics, the difference between the lowest score (i.e., average percent of students passing items) a country received in a topic area and the highest score ranged from 20 to 55 (percent of students passing). That is, a country could have anywhere from 20% to 55% more students passing items, on average, within a topic area. The average difference was 39%. The average of country standard deviations of scores was 10. Differences in science scores ranged from 20 to 51 with an average of 35. The average of country standard deviations of science scores was 7.8.

When looking at ranks, 17 of 42 countries ranked in the top 5 in at least one of the mathematics topic areas and half of the 42 countries ranked in the top 5 in at least one of the science topic areas. Thirty-one of the countries have ranks on mathematics topics that fall in at least three different quartiles, and 30 countries have ranks on science topics that fall in at least three different quartiles. Differences between minimum and maximum ranks within a country average 18 for math (standard deviation of 5) and 23 for science (standard deviation of 6). Country performance at the item level shows even more variability. Twenty-three countries rank first on at least one mathematics item, and 26 countries rank first on at least one science item.

Variation in country-level performance is also seen when looking across items grouped by performance type within topics and also when looking across items grouped by item type. As an illustration of the first point, items testing fractions and testing equations and formulas were divided into three groups: those testing for a basic understanding or knowledge, those testing routine procedures, and those testing problem solving or reasoning. Thus, three sets of scores and ranks were calculated for each topic based upon the type of performance that was expected of students. Differences in ranks within topics ranged from 0 to 20 with an average of 7. Thirteen countries had a difference of 10 ranks or more within each topic. Scores and ranks were also calculated for items within three different item types: multiple choice, short answer, and extended response. Differences in country ranks across these three item types ranged from 0 to 20 with an average difference of 5 ranks in math and 8 ranks in science. Most countries performed better on items with more basic performance expectations or multiple choice item formats, but that was not always the case.

The reduction in variation as scores get further away from individual items is striking. Some of this variation may be explained by measurement error - especially when looking at the item level. However, measurement error most likely does not account for all the variation. Needless to say, the

variation is likely due to differences in curriculum and instruction, and these differences would not be evident through analyses of total scores or even sub-scale scores.

Within-country variations

Another point to keep in mind when searching for curricular effects is that within countries, students do not always receive the same curriculum. In fact, some countries can have as much within-country variation as there is between-country variation. Just as a total achievement score may hide much of the topic-level variations and effects that can be attributable to curriculum, country-level scores may do the same. When looking for curricular effects then, it will be important to match individuals or sub-groups of students to the curriculum to which they were exposed.

Data from the United States can be used to illustrate this point. The United States as a whole ranked below the international mean in mathematics achievement of 13 year-old students. Yet, a group of districts within the United States who participated in TIMSS as a unit not only scored above the mean in mathematics, but also tied Singapore for a ranking of first. The high achievement of such sub-groups of students within the United States is lost in overall-country rankings. Additionally, the state of Minnesota participated in TIMSS also as a unit. They ranked no better than the United States as a whole in mathematics. However, in earth science (a topic that, by consensus, is emphasized by Minnesota teachers) the state of Minnesota tied for a ranking of first. Focusing only on a country-level score and a total score masks these varying patterns of achievement of sub-groups within the U.S. total country mean. As another example, 13 year old students in the U.S typically take one of three mathematics courses (regular, pre-algebra, or algebra) and one of three science courses (earth science, life science, or physical science). Certainly, students within each of these different courses will demonstrate very different patterns of achievement. Other examples exist cross-nationally. Switzerland has cantons with varying curricula and Belgium participated in the

study as two separate systems- a French speaking and a Flemish speaking. Country-level scores would not be able to present the entire picture of variations in achievement in are likely in these situations.

Curriculum-Sensitive Statistics

Researchers analyzing data from cross-national studies not only must decide how to aggregate data in analyses and reporting, they also must decide which statistics are most favorable for demonstrating or picking up on curricular effects. The most popular statistic to use is average performance (of countries or sub-groups, on a total score or sub-scores). However, with mean scores, it is difficult to separate student status, (for example, at eighth grade) with growth (Burstein, 1993). In the TIMSS curriculum analyses, textbook data were collected on countries' math and science curriculum for the upper grade of testing years. Thus, one indicator of curricular effect may be the performance differences between students in the lower grade and students in the upper grade. Ranking of countries on these differences, shows different ordering than for overall achievement (Beaton, Martin, Mullis, Gonzales, Kelly, & Smith, 1996; Beaton, Martin, Mullis, Gonzales, Smith, & Kelly, 1996; Burstein, 1993; Jakwerth, 1996). Other statistics that may be useful are variances or standard deviations, particularly if there is variation in access to opportunity-to-learn within countries. Countries with more equitable distributions may have smaller standard deviations than countries with wide variations. Simply reporting mean performance would hide this variation.

Conclusion

The decisions that researchers make in their analyses have an impact on the type of results they find. It is important the they make decisions that are sensitive to the effects they are seeking to

find. This is why the United States TIMSS is pursuing various avenues of research, including the following

- examination and accounting for curriculum differences within countries while conducting cross-national comparisons,
- examinations of content and performance expectation links across TIMSS across TIMSS data sets,
- examination of matches between the TIMSS achievement test and various measures of the curriculum,
- examination of performance expectations in addition to content-based scores,
- examinations of individual item scores, and
- examination of differences in performance between the upper and lower grades of testing.

In the current period of educational reform, cross-national studies are receiving renewed attention as educational systems across the world strive for "world class" standards and fight to maintain or gain competitive economic footing (Linn & Baker, 1995; Porter, 1990; Schmidt & Valverde, 1995). The results of such studies are useful for both accountability and school improvement. However, researchers and policy-makers cannot allow themselves to be lured into the international horse race and to be swayed by public demands for simplistic results and explanations. The international educational system is varied and complex, and analyses of this system should reflect this complexity.

> My answer to people who want comparative standings is to give them comparative standings - lots of them: in different topics, at different ages, with different kinds of tasks, both unadjusted and adjusted for factors such as national curricula and proportion of students in school. Recognizing that no single index of achievement can tell the full story and that each has its own limitations, we increase our understanding of how nations compare by increasing the breadth of our vision. Even so, however, simply ascertaining nations' relative standing tells us little about how to set educational policy or improve instructional practice.(Mislevy, 1995, p.419)

Validity Issues in Cross-national Relational Analyses:

A Meta-Analytic Approach to Perceived Gender

Differences on Mathematics Learning

I gratefully acknowledge the assistance of Dr. Betsy J. Becker for her comments and editing on a draft of this article, and the TIMSS US National Research Center for the permission of using its data.

Inquiries about this paper should be sent to Wen-Ling Yang, who is now at the Department of Counseling, Educational Psychology, and Special Education, Michigan State University, East Lansing, MI 48824.

Abstract

Introduction. In cross-national research, it may be problematic to analyze data from various countries in one single study when there are distinct country characteristics. In this study, therefore, cross-national data were analyzed in a unique way that important country characteristics are taken into account. Specifically, study outcomes of various countries were treated as independent research results and a meta-analytic approach was applied to synthesize study outcomes across countries while important country features were considered.

Unlike traditional qualitative review methods, which are judgment-based and fail to provide statistical justification for the similarities or differences found among countries, meta-analysis has great potential for improving the validity of cross-national research. Meta-analytic techniques are useful in combining homogeneous country outcomes for the estimation of a cross-countries average outcome. It is also effective in detecting heterogeneity in various country outcomes and offers useful strategies to develop sensible models to explain between-countries differences due to unique country characteristics or different study designs. The applicability and advantages of meta-analysis in summarizing cross-national study outcomes was explored in this study, and the cross-countries outcomes I analyzed was gender difference in students' perceptions about whether girls or boys would do better on mathematics.

Gender difference-- Focus of meta-analysis. Gender plays an important role in students' perceptions or beliefs about mathematics learning (Fennema & Sherman, 1977). It is also found that students' perceptions or beliefs about their own learning correlate with their academic achievement (Schunk, 1981). Gender difference in students' beliefs about gender differences in math learning, therefore, is likely to contribute to the gender difference found in math achievement. It is thus important to study the possible causes of the gender difference in students' beliefs.

Data and Study design. The data analyzed were the results of the field-trial version of TIMSS student questionnaires. The subjects were 7th- and 8th-grade high-school students. Data from 25 countries/regions were available for my study. I first analyzed the data for individual countries using a multiple regression model. Student gender and six other important variables were used as predictors in the multiple regression model. Then I collected various summary statistics from the 25 country-level studies and used them as meta-analyses indicators to synthesize study outcomes across countries.

These meta-analysis indicators included (a) R^2_{total} from the multiple regression model, (b) partial R^2 due to gender, (c) effect size for gender difference, (d) partial regression coefficient for gender ($partial\ \vec{\beta}_{gender}$), and (e) p values. Theoretical strengths and relative merits of these indicators were compared in this study.

Meta-analysis procedures and Illustrative results. Using findings from my study, I illustrated the application of typical meta-analysis procedures for cross-national comparisons. First, sampling bias due to small sample size is corrected for individual country outcomes. Secondly, population variances of country outcomes are estimated. Then, these country outcomes are tested for their homogeneity. If country outcomes appear homogeneous, as the partial R^2_{gender} s of the 25 countries found in my study, one can estimate the cross-counties common parameter ρ^2_{gender} and test its statistical significance.

In the case of the partial R^2_{gender} s, the 95% confidence interval plot showed quite a bit of consistency among the 25 R^2_{gender} s. Homogeneity test statistic Q (df=24, p=.055) further indicated that the population ρ^2_{gender} s were homogeneous across countries. Therefore, a variance-weighting method (Shadish & Haddock, 1994) was used to combine the country outcomes. The average population ρ^2_{gender} was estimated to be .021 (significant for α=.05) with a standard error estimate of .005. This meta-analysis result suggested that, across various countries, after the effects of the other important predictors were controlled, gender still explained a significant amount of variance in students' perceptions of gender differences in learning mathematics. Specifically, girl students thought girl students would do better on math, whereas boy students thought boy students would do better. However, the practical significance of this finding should be addressed because of the small value of the average-parameter estimate.

When country outcomes are heterogeneous, as the $partial\ \vec{\beta}_{gender}$ s in my study, outlier analysis (Hedges & Olkin, 1985) can be conducted to identify extreme cases. Both empirical evidence and judgment are required to determine whether cases are outliers and whether they should be removed from the analysis. In addition, moderator analysis (Eagly & Wood, 1994) can be used to explain between-countries variations. To deal with the heterogeneous $partial\ \vec{\beta}_{gender}$ s, I used two moderators representing important country characteristics to account for between-countries variations. The moderators were the general level of student math achievement and the general level of educational development. Math-achievement-level turned out not to be very useful in explaining between-countries variances, whereas educational-development-level accounted for a small but significant amount of variance in the $partial\ \vec{\beta}_{gender}$ s (about 4.4%). This suggested that gender difference in students' perceptions somewhat depended on the educational development levels of individual countries. However, within either one of the two educational-development groups of countries, country outcomes were found not homogeneous. A large portion of between-countries heterogeneity was not explained by educational-development-level. Educational-development-level was only slight better than math-achievement-level in explaining cross-countries differences in gender difference. If important moderators can be found to model between-countries variances, meta-analysis will show much more analytic strength.

Reasoning inconsistent meta-analysis results. I compared various meta-analysis results and provided explanations and implications for inconsistent results. For example, the meta-analysis results yielded by partial R^2_{gender} s and partial

$\vec{\beta}_{gender}$ s seemed not consistent. While the homogeneity test using partial R^2_{gender} s indicated between-countries homogeneity, the test using partial $\vec{\beta}_{gender}$ s suggested heterogeneity across countries. Possible explanations include (a) $\vec{\beta}_{gender}$ incorporates information on the direction of gender difference, whereas R^2_{gender} doesn't, (b) although the R^2_{gender} s were tested homogeneous, the p value (=.055) of test statistic Q seemed marginal, (c) the estimation for the variances of the partial R^2_{gender} s can be improved, and (d) the sensitivity of homogeneity test to the scales of various statistics needs to be addressed.

One important implication for inconsistent meta-analysis findings is that various statistics may have differential merits for meta-analysis. Probably, this is in part due to the differential approximations of the distributions of different statistics.

Important meta-analysis issues elaborated. Several important meta-analysis issues were elaborated in my paper, including the non-directional nature of R^2, effectiveness of moderators, and the relative merits of various meta-analysis indicators. Suggestions are made for future studies.

Conclusions-- Advantages of meta-analysis. To conclude this paper, an overall evaluation on the applicability and advantages of meta-analysis for cross-national comparisons is provided. As demonstrated in my study, meta-analysis is useful for integrating homogeneous country outcomes, and it is effective in detecting and explaining substantial country differences. In addition, various statistics can be used as meta-analysis indicators and a variety of meta-analytic methods are available for combining cross-countries outcomes. Specifically, country outcomes can be treated as independent studies and be meta-analyzed. Important country features or research design features can be incorporated as moderators to further examine cross-countries differences. Furthermore, if different models are used in different countries, meta-analysis can take into account these model differences while meta-analyzing country outcomes. Therefore, meta-analysis is expected to improve the validity of cross-national comparisons based upon a quantitative approach.

Although I reported gender difference in this paper, it is easy to see that meta-analytic approaches can be used to analyze any other variables of interest within a particular context. For instance, to address the issue of the match between curriculum and test, one can use the degree of the match, or opportunity-to-learn, as a moderator to account for cross-countries differences when student achievement is summarized across countries.

Author's Note.

In addition to meta-analysis, alternative approaches such as hierarchical linear models (HLM) may also be used for cross-countries studies. When many controls are possible and country outcomes are of the same scale, HLM will be feasible. However, in real world, such situation is rare and difficult to achieve. With TIMSS' argument for independent country outcomes, instead of treating country outcomes as data points in one big study, meta-analysis seems more practical. In the future, studies can be done to compare the relative strength and effectiveness of these two approaches in synthesizing cross-national study outcomes.

LIST OF REFERENCES

Airasian, P.W., & Madaus, G.F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*(2), 103-118.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Baker, D.P. (1993, April). Compared to Japan the U.S. is a low achiever...really: New evidence and comments on Westbury. *Educational Researcher, 22*(3), 18-20.

Beaton, A.E., martin, M.O., Mullis, I.V.S., Gonzales, E.J., Kelly, D.L., & Smith, T.A. (1996, Nov.). *Mathematics achievement in the middle school years*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Beaton, A.E., martin, M.O., Mullis, I.V.S., Gonzales, E.J., Smith, T.A., & Kelly, D.L., (1996, Nov.). *Science achievement in the middle school years*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Berliner, D.C. (1993, Fall). International comparisons of student achievement: A false guide for reform. *National Forum*, 25-29.

Bracey, G.W. (1995, Oct.). The fifth Bracey report on the condition of public education. *Phi Delta Kappan*, 149-160.

Burstein, L. (1991). Conceptual considerations in instructionally sensitive assessment. Los Angeles: Center for Research in Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 335367)

Burstein, L. (1993). Studying learning, growth, and instruction cross-nationally: Lessons learned about why and why not engage in cross-national studies. In Burstein, L. (Ed.) The IEA Study of Mathematics III: Student Growth and Classroom Processes. New York: Pergamon Press. (p.xxvi-lii).

Burstein, L., Aschbacher, P., Chen, Z., Lin, L., & Sen, Q. (1990). *Establishing the content validity of tests designed to serve multiple purposes: Bridging secondary-postsecondary mathematics*. Los Angeles, CA: UCLA Graduate School of Education. CSE Dissemination Office.

Crocker, L.M., Miller, M.D., & Franks, E.A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*(2), 179-194.

Crocker, L., Llabre, M., & Miller, M.D. (1988). The generalizability of content validity ratings. *Journal of Educational Measurement, 25*, 287-299.

Cronbach, L.J. & Gleser, G.C. (1953). Assessing similarity between profiles. *The Psychological Bulletin, 50*(6), 456-473.

Cronbach, L.J. (1971). Test validation. In Thorndike, R.L. (Ed.), *Educational Measurement*. Washington, DC: American Council on Education.

Eagly, A. H., & Wood, W. (1994). Using research syntheses to plan future research. In Harris Cooper & Larry V. Hedges (Eds.), *The handbook of research synthesis* (pp.485-500). New York: Russell Sage Foundation.

Fennema, E. L., & Sherman, J. A. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal, 14,* 51-71.

Fitzpatrick, A.R. (1983). The meaning of content validity. *Applied Psychological Measurement, 7*(1), 3-13.

Gamoran, A., Porter, A.C., Smithson, J., & White, P.A. (1996, March). *Upgrading high school math instruction: Improving opportunities for low-achieving, low income youth.* A paper presented at the annual meeting of the American Education Research Association, New York, NY.

Garden, R.A., & Orpwood, G. (1996). Development of the TIMSS achievement tests. In IEA (Eds.), *Third International Mathematics and Science Study Technical Report Volume I: Design and Development.* Boston: Bocton College.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis,17*(3), 323-336.

Guskey, T.R., & Kifer, E.W. (1990). Ranking school districts on the basis of statewide test results: Is it meaningful or misleading? *Educational Measurement: Issues and Practice, 9*(1), 11-16.

Haertel, E. & Calfee, R. (1983, Sum.). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20*(2), 119-132.

Hedges, L. V., & Olkin I. (1985). *Statistical methods for meta-analysis.* Boston: Academic Press.

Husen, T. (1982). *A cross-national perspective on assessing the quality of learning.* Washington, DC: National Commission on Excellence in Education. (ERIC Document Reproduction Service NO. ED225992).

Husen, T. (1983). Are standards in US schools really lagging behind those in other countries? *Phi Delta Kappan, 64,* 455-461.

Husen, T. (1987). Policy impact of IEA research. *Comparative Education Review, 20,* 81-92.

Jakwerth, P.M. (1996). *Evaluating the content validity of cross-national achievement tests.* Unpublished doctoral dissertation, Michigan State University, Michigan.

Kupermintz, H., Ennis, M.M., Hamilton, L.S., Talbert, J.E., Snow, R.E. (1995, Fall). Enhancing the validity and usefulness of large-scale educational assessments: I. NELS: 88 mathematics achievement. *American Educational Research Journal, 32*(3), 523-554.

LaPointe, A.E. (1991). NAEP: A national report card for education and the public. *The Assessment of National Goals: Proceedings of the 1990 ETS Invitational Conference,* 47-62.

Leinhardt, G. (1983). Overlap: Testing whether it is taught. In Madaus, G.F. (Ed.), *The Courts, Validity, and Minimum Competency Testing.* Boston: Kluwer-Nijhoff Publishing.

Leinhardt, G., & Seewald, A.M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18*(2), 85-96.

Linn, R.L. (1987). *State-by-state comparisons of student achievement: The definition of the content domain for assessment.* (Technical report #275). Los Angeles, CA: University of California Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Linn, R.L. (1988). Accountability: The comparison of educational systems and the quality of test results. *Educational Policy, 1,* 181-198.

Linn, R.L., & Baker, E.L. (1995). What do international assessments imply for world-class standards? *Educational Evaluation and Policy Analysis, 17*(4), 405-418.

Maeroff, G. (1991). The public's expectations for assessment of National Educational Goals. *The Assessment of National Goals: Proceedings of the 1990 ETS Invitational Conference,* 87-95.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis,17*(3), 305-322.

Mehrens, W.A. (1984, Fall). National tests and local curriculum: Match or mismatch? *Educational Measurement: Issues and Practice.* 9-15.

Mehrens, W.A., & Lehmann, I.J. (1991). *Measurement and evaluation in education and psychology.* Fort Worth: Holt, Rinehart and Winston, Inc.

Mehrens, W.A., & Phillips, S.E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357-370.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York, New York: American Council on Education, Macmillan Publishing Company.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R.J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis, 17*(4), 419-437.

Muthen, B., Huang, L., Jo, B., Khoo, S., Goff, G., Novak, J., & Shih, J. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis,17*(3), 371-403.

Nitko, A.J. (1989). Designing tests that are integrated with instruction. In R.L. Linn (Ed.), *Educational Measurement* (pp. 453-474). New York, New York: American Council on Education, Macmillan Publishing Company.

Passow, A.H. (1984). The IEA national case study. *Educational Forum, 48,* 469-487.

Phillips, S.E., & Mehrens, W.A. (1988). Effects of curricular differences on achievement test data at item and objective levels. *Applied Measurement in Education, 1*(1), 33-51.

Porter, A.C. (1990). Assessing national goals: Some measurement dilemmas. *The Assessment of National Goals: Proceedings of the 1990 ETS Invitational Conference,* 21-42.

Postlethwaite, T.N. (1987). Comparative educational achievement research: Can it be improved? *Comparative Education Review, 31,* 150-163.

Purves, A.C. (1987). The evolution of the IEA: A memoire. *Comparative Education Review, 31,* 10-28.

Raizen, S.A. & Jones, L.V. (1985). *Indicators of precollege education in science and mathematics: A preliminary review.* Washington, D.C.: National Academy Press.

Robitaille, D.F., McKnight, C., Schmidt, W.H., Britton, E., Raizen, S., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science.* Vancouver, Canada: Pacific Educational Press.

Romberg, T.A., & Wilson, L.D. (1992). Alignment of tests with the Standards. *The Arithmetic Teacher, 40*(1), 18-22.

Schmidt, W.H. (1983). Content biases in achievement tests. *Journal of Educational Measurement,20*(2), 165-178.

Schmidt, W.H., & McKnight, C.C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis,17*(3), 337-353.

Schmidt, W.H., McKnight, C.C., & Raizen, S.A. (1997). *A splintered vision: An investigation of U.S. Science and Mathematics Education.* Dordrecht, the Netherlands: Kluwer Academic Publishers

Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., & Wiley, D.E. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics.* Dordrecht, the Netherlands: Kluwer Academic Publishers.

Schmidt, W.H., Porter, W.H., Schwille, J.R., Floden, R.E., & Freeman, D.J. (1983). Overlap: Testing whether it is taught. In Madaus, G.F. (Ed.), *The Courts, Validity, and Minimum Competency Testing.* Boston: Kluwer-Nijhoff Publishing.

Schmidt, W.H., Raizen, S.A., Britton, E.D., Bianchi, L.J., and Wolfe, R.G. (in press). *Many visions, many aims: A cross-national investigation of curricular intentions in science education.* Dordrecht, the Netherlands: Kluwer Academic Publishers.

Schmidt, W.H., & Valverde, G.A. (1995). *National policy and cross-national research: United States participation in the Third International Mathematics and Science Study.* Manuscript in preparation, East Lansing, MI: Michigan state University, Third International Mathematics and Science Study.

Schunk, D. H. (1981). Modeling and attributional effects on children's achievement: A self-efficacy analysis. *Journal of Educational Psychology, 73*, 93-105.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In Harris Cooper & Larry V. Hedges (Eds.), *The handbook of research synthesis* (pp.261-284). New York: Russell Sage Foundation.

Stedman, L.C. (1994, Oct.). Incomplete explanations: The case of U.S. performance in the international assessments of education. *Educational Researcher, 23*(7), 24-32.

Walker, D. & Schaffarzick, T. (1974). Comparing curricula. *Review of Educational Research, 44*(1), 83-11.

Westbury, I. (1992, June-July). Comparing American and Japanese achievement: Is the United States really a low achiever? *Educational Researcher, 21*(5), 18-24.

Westbury, I. (1993, April). American and Japanese achievement...again. *Educational Researcher, 22*(3), 21-25.

AREA 1997

TM026943

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: VALIDITY IN CROSS-NATIONAL ASSESSMENTS: PITFALLS AND POSSIBILITIES	
Author(s): Jakwerth, Bianchi, Hovang, Schmidt, Valverde, Wolfe, and Yang	
Corporate Source: U.S.-TIMSS National Research Center Michigan State University	Publication Date: March, 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: [signature]	Position: ASSOCIATE DIRECTOR
Printed Name: GILBERT A. VALVERDE	Organization: US-TIMSS RESEARCH CENTER
Address: MICHIGAN STATE UNIVERSITY 457 ERICKSON HALL EAST LANSING, MI 48824-1034	Telephone Number: (517) 353 7755
	Date: 4/8/97

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 21, 1997

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 1997/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

---

[1]If you are an AERA chair or discussant, please save this form for future use.